Multi-review Summarization

David Purdy
Unaffiliated
83 Marion St. #3
Somerville, MA, USA, 02143
Permanent:
P.O. Box 21061
Oklahoma City, OK, USA, 73156

15 March 2002

Paper ID: none provided

Keywords: summarization, multi-document summarization, multi-review summarization, keyterm spanning

Contact Author: David Purdy - dpurdy@post.harvard.edu

Under consideration for other conferences (specify)? none

Abstract

This paper introduces a different way of addressing the problem of multidocument summarization, by focusing on multiple reviews of a specific item, and this focus has some similarity to a common commercial practice of aggregating numerous reviews of a product or service and condensing these reviews to one thumbnail review. In addition, the source code used in this research will be publicly available for further development.

Multi-review Summarization

Paper ID: none provided

Abstract

This paper introduces a different way of addressing the problem of multi-document summarization, by focusing on multiple reviews of a specific item, and this focus has some similarity to a common commercial practice of aggregating numerous reviews of a product or service and condensing these reviews to one thumbnail review. In addition, the source code used in this research will be publicly available for further development.

1 Introduction

The problem of multidocument summarization has thus far found little traction in research or in commercial applications. Principally, this has been due to the difficulty of organizing information from many different viewpoints, timeframes, and biases, and then deciding how to present this information in a useful manner.

Because most projects in multidocument summarization involve indepently authored narratives, typically of a historical event (broadly speaking, news articles are about historical events, albeit very recent events), there are numerous difficulties involved, including:

(1) different authors, which is common in the case of many news articles on a given topic - e.g. coverage of Middle East events could involve local coverage from Israeli and Arab reporters, as well as reporters based at the United Nations Headquarters, or located in Washington, D.C., or any number of other world capitals;

(2) different timelines, which may have significant or little overlap, or may have gaps depending on the author;

(3) different biases of the authors: an author may view the event as grounds for developing some philosophical or political discussion, while others may use it as an example illustrating some such philosophy. Similarly, they may have different political views, experiences, or intended audiences (e.g. mass market news articles which may be short, versus more in-depth pieces intended for a specialized audience);

(4) overlap and disjunction of sources for the reporting: though the articles may be independent, they may derive from the same statements, interviews with the same sources, or even act as responses to other articles or other published interviews (e.g. one side in a conflict may seek media coverage in response to media coverage from another side).

With these concerns, among many others, it's possible to see that presenting a concise summary of news articles or other historical narratives (e.g. biographies), can be taxing for a human, much less the NLP researcher who seeks to automate these tasks.

In addition, finding commercial application for such summaries can be difficult. Often, a set of news articles on the same topic and from the same time period cover much of the general background, making it redundant to have a separate summary which repeats this material. It can be argued that because this background information is redundant it should be excerpted and presented as a "common knowledge" summary, to be read before reading specific articles; but, then one is left with prioritizing and editing articles based on the presumption of the knowledge of the common background. It becomes a bit of a mess. Finding existing human analogs for this kind of news processing is difficult and rare. Usually such work is done on an infrequent and rare basis, and often by specialists who know the topic very well.

One area with real commercial applications for multidocument summarization involves aggregating and condensing numerous reviews. Perhaps the most notable example is the Zagat Survey series of restaurant reviews, available in most large cities. In such a guide, a restaurant may have been reviewed by anywhere from a dozen up to thousands of people. Similarly, at some college campuses, it is common to have publications which condense student reviews of courses and professors. Now, there is a flood of sources of such reviews, with major online sites such as Amazon (amazon.com), the Internet Movie Database (IMDB) (imdb.com), Epinions (epinions.com), all offering customers the opportunity to review books, electronics, movies, music, stores, services, and more. Some very popular books on Amazon have hundreds of reviews, while Guinness Stout beer, for instance, has over 260 reviews on Epinions, and a review of the film "The Matrix" had over 900 reviews on Epinions and over 2400 on IMDB. For such a situation, it is difficult for any reader (or even a paid editor) to wade through so many reviews. Most sites allow a reviewer to also give a numerical rating to the reviewed item, and then offer a reader an average of the numerical ratings. This numerical average is the best that is presently done automatically in assessing an average review. Such data is frequently used in collaborative filtering applications and services. A notable example is Amazon's recommendation system, which recommends other products based on comparing a user's rating an d purchasing behavior with users who have similar habits.

I address the issue of multi-review summarization starting from a very simple basis. The essential idea is that whatever is being reviewed is likely a noun, and we should first span the important nouns (as a proxy for spanning terms), and then span the adjectives used in describing these nouns. This research work represents a first approach to multi-review summarization, and while the results appear quite good, it is definitely possible to refine this approach for broader use as well as make applications which are customized for individual users.

2 Definitions

Keyterms

For this paper, the keyterms are the N most frequently occurring nouns.

Modifiers

These are the M most frequently occuring adjectives.

Spanning

In the context of summarization, spanning involves taking a small set of fundamental units and seeking to find an optimal selection of sentences (or, more broadly, text passages) which cover (or span) these fundamental units. An example of spanning is to identify the most relevant keyterms of a document, and then select the fewest possible sentences which mention these keyterms. More refined spanning, however, would assign a score to each such selection of sentences, and then find a minimal set of sentences with the best overall score (which may be independent of the scores of individual sentences), thus optimizing the overall relevance of the set.

Spanning allows us to address several issues at once: relevance and redundancy. We assure relevance because we value sentences which mention keyterms from the document. Redundancy is minimized because once a keyterm is mentioned, we seek to span other keyterms, and not repeat a term which has already been covered.

3 Example

For the purposes of testing these algorithms, I used Amazon.com customer reviews for several popular books.[1] It is important to note that licensing constraints did not allow for the luxury of a large corpus of such reviews, so I could not include *tf.idf* filtering, for instance.

A typical review averages about 200 words in length, and a popular book may have upwards of 100 reviews, and some books studied had over

[1] Note that such use is permitted under the fair use doctrine of U.S. copyright law.

600 reviews. As a result, it becomes very difficult for a reader to process so many reviews in a useful manner.

Having obtained these reviews, I made use of several publicly available tools to process these documents. These were: Ratnaparkhi's MXTERMINATOR program for finding sentence boundaries, MacIntyre's script for tokenization, and Brill's rule-based part of speech tagger.

For each aggregate of reviews, the **keyterms** selected were the N most frequent nouns, as identified by the **NN** tag from the Brill tagger. The **modifiers** selected were determined by the **JJ** tag.

For each noun, N_i, its score $S(Ni)$ was equal to the number of total occurrences among all reviews. Similarly, for each adjective A_j, its score $S(Aj)$ was equal to the total number of occurrences in all reviews. All other words were assigned a score of 0. Though this is a very simple measure, subject to many refinements, it still yielded very useful summaries.

Every sentence, composed of words $w_1, ..., w_k$, the score of the sentence is the sum of the score of each word:

$$S_{sentence} = \sum_{i=1}^{i=k} S(w_i) \quad (1)$$

For instance, 115 reviews of Sylvia Nasar's "A Beautiful Mind", a biography of the mathematician John Nash, yields the following top 10 key terms:

count	token	POS tag	rank
331	nash	NNP	1
262	book	NN	2
113	nasar	NNP	3
99	john	NNP	4
88	life	NN	5
75	story	NN	6
71	genius	NN	7
55	movie	NN	8
53	biography	NN	9
51	world	NN	10

Naturally, one would not be surprised to find (1)the author's name [#3 - Nasar], (2) the name of the book's subject [#1 - nash, #4 - john], and various characteristics describing the story: life story [#5 and #6] and biography [#9], and genius [#7]. In addition, a connection to the recent movie is also frequently noted [#8, mentioned 55 times total among 115 reviews].

Adjectives that are frequently used include:

count	token	POS tag	rank
52	mental	JJ	1
45	interesting	JJ	2
43	mathematical	JJ	3
43	great	JJ	4
39	many	JJ	5
38	more	JJR	6
36	other	JJ	7
36	beautiful	JJ	8
35	such	JJ	9
29	own	JJ	10

These are less uniquely applicable to this book, though the first 4 may be rather unique. Nash's mental abilities as well as his psychoses are very central to the book, thus the frequent appearance of "mental" is understandable. However, adjectives such as "many", "more", and "other" are not particularly relevant to this book. Such adjectives may be weighted downward by way of *tf.idf* measures. One may also note that the frequency distribution is much smoother than that for nouns.

With only one exception – the removal of "book"[2] from the keyword list – we get the following top 5 scoring sentences when scoring all the sentences found among all of the reviews, using formula (1):

> All in all, this is a well written and interesting story, with a smattering of interesting background information on mathematics and economics in the 1950s, the psychology of schizophre-

[2] With a large corpus of such reviews, we could use *tf.idf* scoring to reduce the significance of the media type (e.g. book, video, etc.), or other inherent attributes (e.g. restaurant, movie, play, musical). However, the constraints of using Amazon.com prohibited the acquisition of such a sufficiently large corpus.

> nia, and the struggles of Nash, Alicia and other people to deal with mental illness and cope with life. [score: 468]
>
> In short I was charmed by the book, it gave me a lot a material with which to consider the nature of genius, mathematical accomplishment, mental illness and particularly the effect of other people on ones sense of self and what is meant by a whole life. [score: 441]
>
> She tells a very important story that captures the organizational culture of math departments throughout colleges and universities across the country, homosexuality during the McCarthy era, mental illness and the recovery of mental illness, relationships and the importance of them, as well as mathematical theorems - how they developed and the use of them. [score: 418]
>
> It covers a number of interesting subjects: mental illness, mathematics and economics, life in the academic world, "behind the scenes" politics in the Nobel committee, and obviously the complex personality of John Forbes Nash Jr. [score: 413]
>
> However my interest wained as the story progressed and his personal life became more of the focus although it must be said, such a lengthy and detailed description was neccesary in many respects, particulary in nurturing the readers emotional attachment to Nash and his wife Alica. [score: 346]
>
> [Amazon, 2002]

A review using these five sentences would be only slightly longer (about 12% more words) than the typical summary, but is assured of spanning a great many topics. Thus, the reader can quickly assess the utility of reading more reviews of this book or whether to move on to other books.

4 Further Work

There are a number of things that can be developed in further research:

1. Better weightings for nouns and adjectives when scoring a candidate summary. Presently, the frequency counts for both are used, but I believe that one should span on the nouns, then weight adjectives according to some context-based relevance, such as proximity to major nouns.

2. A slightly more sophisticated system would differentiate between the various numerical ratings that reviewers give and produce a summary review for each level. More simply, one could also produce a system which separately summarizes favorable reviews and unfavorable reviews, so that a reader can understand the high and low points of the reviewed matter.

3. In separate testing, based on reviews of electronics equipment, I found that technical terms frequently added noise to a summary review. Since there is often a much higher profit margin, and a much lower general familiarity with electronics products, separate research into how to make good summary reviews for electronics could pay substantial dividends.

4. At some point, merging this work with collaborative filtering work, to produce summaries of reviews by people with similar (or very dissimilar) behaviors could yield a higher degree of customization.

5. A baseline methodology for evaluation needs to be developed. It could be based on user surveys, which would have more immediate application, but research purposes would need more objective reusable metrics. I am not aware of any good method for developing these, but admit that it would be worth pursuing.

5 Conclusion

I believe this work represents new ground in applications and research for multidocument summarization. The existence of companies which hire people to summarize multiple documents indicates that this could be commercially useful, and the fact that outlets for people to submit many reviews online means that it would be difficult for reviewers to keep up in real-time, indicates that it will become necessary for some level of automation of this field.

I believe that the next steps will need to involve corporations which amass large databases of reviews, and the development of new services which make use of these reviews. I also think that development of evaluation methodologies should be tied directly to the purposes of the application.

As the source code for these algorithms and eventual improvements will be released periodically, I invite others to contribute their thoughts and improvements.

References

Amazon.com customer reviews for Sylvia Nasar's "A Beautiful Mind". 6 March 2002. http://www.amazon.com

Brill, Eric. August 12, 1994. Rule-based Part of Speech Tagger.
http://www.cs.jhu.edu/~brill/RBT1_14.tar.Z

Brill, Eric. Some advances in rule-based part of speech tagging. Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94), Seattle, Wa., 1994.

MacIntyre, Robert. 1995. Sed script to produce Treebank tokenization on arbitrary raw text. http://www.cs.columbia.edu/~eeskin/projstud/erics/tokenizer.sed

Mani, Inderjeet. 2001. Automatic Text Summarization. Philadelphia: John Benjamins Publishing Company.

Mani, Inderjeet and Maybury, Mark T. 1999. Advances in Automatic Text Summarization. Cambridge: MIT Press.

Ratnaparkhi, Adwait. 1997. MXTERMINATOR - sentence boundary detector. ftp://ftp.cis.upenn.edu/pub/adwait/jmx/